A Methodology for Quantifying the Effect of Missing Data on Decision Quality in Classification Problems

Feldman, Michael ; Even, Adir ; Parmet, Yisrael

Abstract: Decision-making is often supported by decision models. This study suggests that the negative impact of poor data quality (DQ) on decision making is often mediated by biased model estimation. To highlight this perspective, we develop an analytical framework that links three quality levels – data, model, and decision. The general framework is first developed at a high-level, and then extended further toward understanding the effect of incomplete datasets on Linear Discriminant Analysis (LDA) classifiers. The interplay between the three quality levels is evaluated analytically - initially for a one-dimensional case, and then for multiple dimensions. The impact is then further analyzed through several simulative experiments with artificial and real-world datasets. The experiment results support the analytical development and reveal nearly-exponential decline in the decision error as the completeness level increases. To conclude, we discuss the framework and the empirical findings, elaborate on the implications of our model on the data quality management, and the use of data for decision-models estimation.

DOI: https://doi.org/10.1080/03610926.2016.1277752

A Methodology for Quantifying the Effect of Missing Data on Decision Quality in Classification Problems

Michael Feldman[1], Adir Even[2], Yisrael Parmet[2]

[1]Department of Informatics, University of Zurich, Zurich, Switzerland

[2]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheba, Israel

ABSTRACT

Decision-making is often supported by decision models. This study suggests that the negative impact of poor data quality (DQ) on decision making is often mediated by biased model estimation. To highlight this perspective, we develop an analytical framework that links three quality levels – data, model, and decision. The general framework is first developed at a high-level, and then extended further toward understanding the effect of incomplete datasets on Linear Discriminant Analysis (LDA) classifiers. The interplay between the three quality levels is evaluated analytically - initially for a one-dimensional case, and then for multiple dimensions. The impact is then further analyzed through several simulative experiments with artificial and real-world datasets. The experiment results support the analytical development and reveal nearly-exponential decline in the decision error as the completeness level increases. To conclude, we

discuss the framework and the empirical findings, elaborate on the implications of our model on the data quality management, and the use of data for decision-models estimation.

Keywords: Data quality, Model quality, Decision quality, Completeness, Linear Discriminant Analysis (LDA)

1. INTRODUCTION

The common phrase "Garbage in Garbage Out" reflects a major concern in the field of information systems (IS) – the negative impact of data quality (DQ) defects on decision making. This study offers a novel perspective for understanding this impact by exploring the mediating role of decision models. Decision-making is often supported by models (e.g., theoretical, analytical, visual, statistical) – representation that reflects real-world phenomena or behaviors. Decision models permit prediction of future behavior and, by that, support decisions and actions. This study focuses on analytical models that represent decision problems as parameterized mathematical formulations. It suggests that the negative impact of poor DQ on decision making is often mediated by biased estimation of decision-model parameters. To support this argument, this study offers both analytical and experimental evaluation of this impact.

The need to managed DQ, and the associated challenges and costs, have long been recognized and discussed in information systems (IS) research and practice. The challenges and costs have magnified immensely with the rapid growth in the volumes and the variety of data resources collected in organization. This growth motivates the questions that underlie this study – to what extent would DQ degradation harm decision making? Would the improvement in decision making justify the cost of preventing or correcting DQ defects? To shed light on the link between DQ and decision making and to better understand the mechanisms behind it, this study takes a novel perspective that has not been addressed much in DQ research. It observes the biased estimation of decision-model parameters due to flawed data, and the associated degradation in decision performance. At a high-level, this association can be well understood – high-quality data forms better models which, in turn, promote better decisions. However, in this study we

wish to make progress beyond the high-level understanding of models' estimation impact - this by developing an analytical framework that quantifies the effects and by evaluating it empirically through simulated experiments. The framework (Figure 1) distinguishes between three stages of a model-driven decision process: data, model, and decision. Accordingly, it defines three levels of quality, each reflecting the corresponding stage: data quality (DQ), model quality (MQ), and decision quality (annotated by CQ, as this study focuses on classification decisions).

Data: real-world entities, relations, actions, and the associated properties can be abstracted into data through a process of acquisition. Data records, which are often subject to DQ defects, are grouped into datasets and stored electronically in databases. This study focuses on completeness – a DQ dimension that reflects the extent of missing records in datasets and/or missing attribute values in records.

Model: this study addresses data-driven models – analytical formulations of real-world behaviors, for which the parameters are estimated from data samples. MQ refer assessment of model goodness, i.e., the extent to which the model reflects the real state of the world in a reliable manner. When DQ degrades - e.g., lower completeness due to missing data - MQ is likely to degrade too, as the estimation of model parameters might become less reliable. This study focuses on classification models and, more specifically, on the commonly used binary Linear Discriminant Analysis (LDA) classifier. Training a model as such requires a dataset of real-world samples, where an incomplete dataset is likely to result in lower MQ.

Decision: with the focus on classifiers, CQ is defined as the extent to which classification decisions are correct. Obviously, it is affected by MQ, as flawed model are likely to lead to wrong decisions. With classifiers, CQ is commonly assessed in terms of assigning an instance to

the correct class. With classifiers that are trained with an incomplete dataset of samples (lower DQ and MQ), the likelihood of assigning an instance to the wrong class is higher (lower CQ).

The framework shows explicit and quantifiable impacts between DQ, MQ, and CQ. Further, it links quality degradation to cost, toward assessment of cost-benefit tradeoffs and optimization of DQ management policies accordingly. The next section provides theoretical background and introduces the analytical development of the proposed methodology. The methodology is then evaluated empirically through simulated experiments with both artificial and real-world data. The concluding section underscores the study's key contributions and implications for practice. It also discusses its limitations and proposes directions for future research.

2. BACKGROUND AND FRAMEWORK DEVELOPMENT

This section reviews relevant literature that influenced the development and evaluation of our research framework for assessing the impact of DQ on MQ and CQ (Figure 2). The framework reflects scenarios in which previously collected data instances (the "training set") are used to estimate a model (a classifier, in this study). The model is then used to classify a new instance, not previously included in the training set. The section is organized along the three framework levels - data, model and decision – and discusses the associated quality aspects, where DQ reflects the goodness of the training set, MQ reflects the goodness of the estimated model, and CQ reflects the goodness of classification decisions that are based on that model. Following a general definition of DQ, MQ, and CQ, we extend the discussion further for the specific case of the binary Linear Discriminant Analysis (LDA) classifiers.

2.1 Data Quality

The need to collect, store and process data is broadly recognized and addressed in the field of information systems (IS) - and so is the need to manage data at a high quality level. Data abstracts real-world entities, relations and actions. High-quality data implies abstraction that reflects the true real-world state well. The consequences of poor data quality are experienced across the board – from damages to ongoing operational tasks and processes, to negative impact at the strategic level on firms' profitability (Madnick et al., 2009). A plethora of studies have pointed out the negative impact of low DQ on decision making and decision support systems (e.g., Batini et al., 2009; Even et al., 2010).

DQ is usually assessed along different dimensions (e.g., accuracy, currency, validity, and reliability), each reflecting a different type of DQ hazards (Pipino et al., 2002; Even & Shankaranarayanan, 2007). Completeness reflects the extent to which items are missing in data collections. Missing values may occur due to unavailability of data at time of acquisition, reluctance to provide data due to privacy concerns, flawed data collection procedures, data processing mistakes, and possibly other reasons (Redman, 1996; Even et al., 2007). Missing values may severe implications for data usability and potential contribution – it might harm decisions based on data analysis (Ballou and Pazer, 2003; Even et al., 2010), and bias the outcomes of machine learning and data mining algorithms (Luengo et al., 2012). A plethora of studies (e.g., García-Laencina et al., 2007; Luengo et al., 2010/2012) have suggested and evaluated methods for data imputation – the filling of missing values with estimated ones, or overcoming biases resulted by data incompleteness. Rubin (1976) discusses three possible mechanisms for the formation of missing values, each reflecting a different form of missing-data

probabilities and relationships between the measured variables, and each may lead to different imputation methods (Luengo et al., 2012):

- **Missing Completely at Random (MCAR):** a missing value that cannot be related to the value itself or to other variable values in that record. This is a completely unsystematic missing pattern and therefore the observed data can be thought of as a random unbiased sample of a complete dataset.
- **Missing at Random (MAR):** cases in which a missing value is related to other variable values in that record, but not to the value itself (e.g., a person with a "marital status" value "single", has a missing value in the "spouse name" attribute). In other words, in MAR scenarios, incomplete data can be partially explained and the actual value can be possibly predicted by other variable values.
- **Not Missing at Random (NMAR):** the missing value is not random and depends on the actual value itself; hence, cannot be explained by other values (e.g., an overweight person is reluctant to provide the "weight" value in a survey). NMAR scenarios are the most difficult to analyze and handle, as the missing data cannot be associated with other data items that are available in the dataset.

The framework developed in this study assumes an MCAR pattern at the record level – each data record is either available or missing entirely, and the likelihood of a record to be missing does not depend on other records, or the actual values in that missing record. This pattern is equivalent to randomly deleting dataset records, or to taking a random sample rather than analyzing the entire population.

Following common terminology (Duda et al., 2012), we refer to the model estimation as "training" and to the dataset $\{(X, Y)_n\}$ used for it as a "training set". The annotation reflects N records (indexed $1..N$), where X is a vector of M attributes (indexed $1.. M$), each reflecting a certain property of a real-world instance. The Y component is a $1..K$ integer that associates the record with one among K classes. Following common DQ measurement schemas (Even & Shankaranarayanan; 2007), each record is associated with a Q_n measurement of completeness - 0, if one or more attribute values (or the entire record) are missing (i.e., NULL), and 1 is the record is complete. The quality of the entire dataset Q^D, in terms of completeness, is defined as the rate of non-missing values, where $Q^D=1$ reflects a complete training set:

$$Q^D = \frac{1}{N}\sum_{n=1}^{N} Q_n, \quad 0 \le Q^D \le 1 \tag{1}$$

2.2 Model Quality

The use of mathematical models for decision support is often termed as "decision calculus" or "management science". It is a common practice in a broad range of contexts such as accounting, financial, marketing, production, services, and many others. Literature offers a plethora of models for decision support and analysis, applying a variety of techniques such as optimization, dynamic programming, stochastic modeling, simulation, statistics, and classification models (Bhargava et al., 2007). Such models are often embedded in decision support systems (DSS) – software-based utilities that help decision-makers in analysis and prediction. Model-based DSS typically consists of three stages (Shim et al., 2002): 1) Model generation in a form acceptable to

the model solver 2) Algorithmic solution of the model, and 3) Analysis (e.g., 'what-if' investigations) and interpretation of the solution.

In classification decisions a certain object is associated with one category (or class) among a set of choices. Given two or more populations and a set of associated variables, the desire in classifications is to define a set of rules that maximizes the separation among the groups by minimizing either the total misclassification probability or the average misclassification cost (Van Der Heijden et al., 2005). Classification models (a.k.a. classification algorithms or classifiers) apply rules for allocating or assigning observations to groups. A classification algorithm can be based either on supervised or on unsupervised learning. In supervised learning, category labels are known in advance for the training set and the algorithm aims to reduce the misclassification costs within the training set. In unsupervised learning, training-set data is unlabeled and the algorithm forms classes based on degree of similarity among observations (Duda et al., 2012).

Literature has identified a variety of classifier designs, which can be categorizes at a high-level (Hastie et al., 2005) along: ***A) Nearest-Neighbor (N-N)*** classifiers assign an object to a class, based on the similarity to training-set samples. The assessment of similarity is based on a certain appropriate metric (often termed as distance), which can be used for classifying patterns by template matching or by minimum-distance criteria. ***B) Probabilistic*** classification methods are based on a analyzing statistical properties of the different classes. The Bayesian classifier, for example, assigns a pattern to the class with the maximum posterior probability. It can be modified to take into account the costs associated with different types of misclassifications.

Maximum Likelihood decision rules attempt at minimizing the sum of the error probabilities without dependency on prior knowledge, and ***C) Error-Minimization*** classifiers consider a certain criterion – e.g., the Mean-Squared Error (MSE) between the classifier output and some preset target value.

The Linear Discriminant Analysis (LDA) classifier belongs to the latter category. It separates between classes by constructing discriminants, where in a case of equal covariance matrixes of class distributions the classes are separated by linear hyper-plains (Duda et al., 2012; Webb, 2003). The LDA assigns an input vector X to either class Y_0 (referred to as "negative", with no loss of generality), or to class Y_1 (referred to as "positive"). LDA assumes that the two classes reflect normally-distributed populations with the same covariance matrix $\sum$ - each with prior probabilities V_0 and V_1 respectively (that sum up to *1*), and different means μ_0 and μ_1 respectively. The LDA classifies vector X of continuous attributes to Y_0 or Y_1 by minimizing the probability of misclassification by assessing the log-ratio of the common probabilities. The prediction leads either to Y_1, if the result greater than 1 or to Y_0 otherwise. The expansion of the log-ratio while the quadratic component (i.e., $X^T \Sigma X$) dropped due to covariance similarity assumption leads to linear decision boundary, where $W = \Sigma^{-1}(\mu_1 - \mu_0)$ and $W_0 = -1/2\ \mu_1^T \Sigma\ \mu_1 + 1/2\ \mu_0^T \Sigma\ \mu_0 + \log\big(V_0(1 - V_0)\big)$:

$$Log\left(\frac{P(X, Y = Y_1)}{P(X, Y = Y_0)}\right) = W^T X + W_0 \tag{2}$$

Figure 3a depicts a scalar ("1 dimensional") binary LDA classifier, in which case the classification rule can be simplified to: X is classified as Y_1 if $X > A$, or classified as Y_0 otherwise. Figure 3b shows a 2-dimensional binary LDA classifier. As seen in both examples, the binary LDA is not a perfect classifier – some misclassifications may occur, as the two populations overlap to an extent. However, given the parameters of the two distributions – the LDA classifier defines the optimal linear separation in terms of minimizing the likelihood of error. To demonstrate our concept, we further develop the scalar (1-dimensionl) case.

This study observes a classifier for which the class-separation parameters are estimated from a training set. The Confidence Interval (CI) is a common approach for assessing the reliability of estimated model parameters (McLachlan, 2004). For example, when estimating a parameter A - CI assessment would allow us to state that "with a confidence of g%, the true value of A resides within the CI of $[\hat{a} - \Delta_1,\ \hat{a} + \Delta_2]$", where $\hat{a}$ is the estimated value. A smaller CI implies a more reliable model, and with classification models, it can be shown that the CI gets smaller with a larger training set (a greater N). Following the CI-assessment concept - we take L, the length of the confidence interval as a measure for model quality - i.e., if the confidence interval is defined by$[\hat{a} - \Delta_1,\ \hat{a} + \Delta_2]$, then $L = \Delta_1 + \Delta_2$. The model-quality metric has to be defined for each model parameter. It has to consider the desired target confidence level 1-ρ, the number of samples N in the complete dataset, and the missing value rate (as reflected by Q^D):

$$Q_A{}^M\left(1-\rho, N, Q^D\right) = L_A\left(1-\rho, N * Q^D\right) \qquad (3)$$

Where

A - The parameter under evaluation

$1-\rho$ - The target confidence level

N - The number of samples in the complete training dataset

Q^D - The data quality level (i.e., the rate of non-missing values)

L_A - The CI for A, given a target confidence level and the sample size

The parameters of two distributions we aim to classify (μ_1, μ_0, and σ) have to be estimated from a "training set" –$\hat{\mu}_1$, $\hat{\mu}_0$, and $\hat{\sigma}$, respectively. With no missing values, the set has N samples per class (a total of $2N$). However, we assume now that some values are missing completely at random (MCAR) from that training set ($Q^D<1$) – i.e., incompleteness distributes evenly between the two classes, and the training set contains Q^D*N samples per class. We annotate the "positive" and "negative" training sets with the missing values by $\{x^1_n\}$ and $\{x^0_a\}$, respectively (in both classes the index *[n]* goes between *1..*Q^DN). Under the MCAR assumption, the following unbiased estimators for the means and the variance can be used (Rubin, 1976):

$$\hat{\mu}_1 = \frac{\sum_{n=1}^{Q^D N} x_n^1}{Q^D * N}, \quad \hat{\mu}_0 = \frac{\sum_{n=1}^{Q^D N} x_n^0}{Q^D * N}$$

$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{n=1}^{Q^D N} \left(x_n^1 - \hat{\mu}_1\right)^2}{Q^D * N - 1}}, \quad \hat{\sigma}_0 = \sqrt{\frac{\sum_{n=1}^{Q^D N} \left(x_n^0 - \hat{\mu}_0\right)^2}{Q^D * N - 1}} \tag{4}$$

$$\hat{\sigma} = \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}{2}} \quad = \frac{\sqrt{\sum_{n=1}^{Q^D N} \left(x_n^1 - \hat{\mu}_1\right)^2 + \sum_{n=1}^{Q^D N} \left(x_n^0 - \hat{\mu}_0\right)^2}}{\sqrt{2\left(Q^D * N - 1\right)}}$$

If distribution parameters are known and under the symmetric case *(V_0=V_1=1/2)*, the optimal classification threshold can be calculated by A= (μ_0+μ_1)/2. As the samples in the training set are drawn from Normally-distributed populations, the estimator $\hat{A}$ is also a normally-distributed random variable, for which we can calculate the expected value $E[\hat{A}]$, and the variance $VAR[\hat{A}]$:

$$\hat{A} = \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}, \qquad E[\hat{A}] = E\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\mu_1 + \mu_0}{2}$$
$$\sigma^2_{\hat{A}} = VAR\left[\hat{A}\right] = VAR\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\sigma^2}{2Q^D * N} \qquad (5)$$
$$\text{hence,} \quad \hat{\sigma}^2_{\hat{A}} = \frac{\hat{\sigma}^2}{2Q^D * N}$$

The rate of missing values (as reflected by data quality measurement Q^D) may directly affect the classification rule, by increasing uncertainty about best classification threshold. As seen in Eq. 5, missing values that follow the MCAR, do not result in bias of expected threshold (the expression $E[\hat{A}]$ does not depend on the data quality level Q^D). However, missing values might affect the confidence level and hence, the model quality Q^M. The estimation variance $VAR[\hat{A}]$ and the associated confidence interval (CI), increase with a higher rate of missing values (lower Q^D). As the estimator for the threshold parameter has a Normal distribution, we can define the confidence interval CI_A for the estimator $\hat{A}$, where the expression $t_{1-\rho/2,\ 2Q^D N-2}$ reflects the 1-ρ quantile of Student-t distribution with ${}_{2Q^D N-2}$ degrees of freedom:

$$CI_A(\rho, Q^D, N) = \begin{bmatrix} \hat{A} - t_{1-\rho/2,\lfloor 2Q^D N \rfloor - 2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}}, \\ \hat{A} + t_{1-\rho/2,\lfloor 2Q^D N \rfloor - 2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}} \end{bmatrix} \quad (6)$$

Accordingly, we calculate the CI-length (the MQ metric, Eq. 4) for the LDA threshold *A*, given a desired confidence level *1–ρ*, *N* samples in the complete dataset, and a DQ level of Q^D:

$$Q_A{}^M(\rho, N, Q^D) = L_A(\rho, N, Q^D) = 2 * t_{1-\rho/2,\lfloor 2Q^D N \rfloor - 2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}} =$$

$$t_{1-\rho/2,\lfloor 2Q^D N \rfloor - 2} * \frac{\sqrt{\sum_{n=1}^{Q^D N}\left(x_n^1 - \hat{\mu}_1\right)^2 + \sum_{n=1}^{Q^D N}\left(x_n^0 - \hat{\mu}_0\right)^2}}{\sqrt{Q^D * N * \left(Q^D * N - 1\right)}} \quad (7)$$

It can be shown that with a large *N*, the Student-*t* distribution can be approximated with a Normal distribution - e.g., with 30 or more degrees of freedom, the error of approximating the probability density function (PDF) of a Student-*t* distribution with a Normal distribution is less than 0.005. Accordingly, the CI-length will be $L_A(\rho, N, Q^D) = 2Z_{1-\rho/2} * \hat{\sigma}_{\hat{A}}$, where Z reflects normal distribution.

2.3 Decision Quality

Decision quality can be assessed in terms of how well the decision reflects an understanding of a prediction of the true world state, and how well the actions taken fit the decision-maker's goals given the true world state. The higher is the quality of data, the more precise it the decision

model and, hence, the higher is the likelihood of making good decisions. With binary classifiers (i.e., K=2), in which the output is either positive (Y=1) or Negative (Y=0), it is common to assess classification performance with the 2-way confusion matrix (Table I). The total number of instance per quadrant (N_{TP}, N_{FP}, N_{FN}, N_{TN}, respectively, where $N_{TP}+N_{FP}+N_{FN}+N_{TN} = N$), are commonly used for assessing classification quality metrics, such as:

- **Accuracy, $Q^{C/A}$** = $(N_{TP} + N_{TN}) / N$**,** the rate of items classified correctly
- **Precision $Q^{C/P}$** = $N_{TP} / (N_{TP} + N_{FP})$, correctness within positive results
- **Sensitivity $Q^{C/S}$** = $N_{TP} / (N_{TP} + N_{FN})$, the ability to detect positive results
- **Specificity $Q^{C/F}$** = $N_{TN} / (N_{TN} + N_{FP})$, the ability to detect negative results

We now develop further the "Classification Accuracy" measurement toward defining a decision quality metric for a binary LDA classifier – classes Y_1 (“positive”) and Y_0 (“negative”), with a-priory probabilities of V_1=V_0=0.5. Each class reflects a Normally-distributed population with different means $\mu_1>\mu_0$ but the same standard deviation σ. With some probability W_{TP} a “positive” item can be classified correctly as “positive”, and with some probability W_{FN}=1–W_{TP} as “negative” ($W_{TP}+W_{FN}$=1). Similarly, with some probability W_{TN} a “negative” item can be classified correctly as “negative”, and with some probability W_{FP}=1–W_{TN} as “positive”. We also assume some known cost U for “False Positive” or “False Negative" misclassifications cases. The LDA model, in that case, has one parameter only – the threshold A that defines the classification rule. A new instance X, with unknown classification, is classified as “positive” if $X>A$, or “negative” otherwise. Based on the assumptions above, it can be shown that with known

distribution parameters (μ_1, μ_0, and σ), the optimal threshold value, in terms of maximizing classification accuracy, is A= (μ_0+μ_1)/2, with a confidence interval of CI_A=0 (as the distribution parameters are known, and not estimated). The probabilities of correct classifications versus misclassification can be calculated accordingly as follows, where Φ is the cumulative normal distribution:

$$W_{TP} = 1 - \Phi\left((A - \mu_1)/\sigma\right) = 1 - \Phi\left(\left(\frac{\mu_0 + \mu_1}{2} - \mu_1\right)\Big/\sigma\right) =$$
$$1 - \Phi\left((\mu_0 - \mu_1)/2\sigma\right) = \Phi\left((\mu_1 - \mu_0)/2\sigma\right)$$

$$W_{FP} = 1 - W_{TP} = 1 - \Phi\left((\mu_1 - \mu_0)/2\sigma\right) = \Phi\left((\mu_0 - \mu_1)/2\sigma\right) \quad (8)$$

$$Due\ to\ symmetry: W_{TN} = W_{TP} = \Phi\left((\mu_1 - \mu_0)/2\sigma\right),$$
$$W_{FN} = W_{FP} = \Phi\left((\mu_0 - \mu_1)/2\sigma\right)$$

The expected decision quality Q^{c*} for this case is:

$$Q^{C*} = Q^C\left(A = (\mu_0 + \mu_1)/2\right) = V_1 * W_{TP} + V_0 * W_{TN} = \Phi\left((\mu_1 - \mu_0)/2\sigma\right) \quad (9)$$

With known distribution parameters (μ_1, μ_0, and σ), the expression in Eq. 9 would be the best possible decision quality that can be obtained. With $\mu_1 - \mu_0 \rightarrow 0$, and/or with $\sigma \rightarrow \infty$, $Q^{C*} \rightarrow 0.5$ (a random "flip of a coin"). With $\mu_1 >> \mu_0$, and/or with $\sigma \rightarrow 0$, $Q^{C*} \rightarrow 1$. The expected decision cost, in that case, would be:

$$C^{C^*} = U * \left(1 - \Phi\left(\left(\mu_1 - \mu_0\right)/2\sigma\right)\right) = U * \Phi\left(\left(\mu_0 - \mu_1\right)/2\sigma\right) \quad (10)$$

Again, with known distribution parameters, this would be the lowest possible decision cost (hence, C^{C^*}). With $\mu_1 - \mu_0 \rightarrow 0$, and/or with very large σ, $C^{C^*} \rightarrow 0.5U$. With $\mu_1 >> \mu_0$, and/or with $\sigma \rightarrow 0$, $C^{C^*} \rightarrow 0$.

We next show the impact of DQ and MQ on the decision quality CQ. With no value for correct classification, but rather some negative cost U for misclassification, we assess decision quality in terms of lowering cost. Eq. 10 set an upper-bound to the cost, when distribution parameters are known. When the parameters are estimated from a sample the decision quality degrades further with a smaller sample size and with lower DQ level. Given a certain threshold $\hat{A}$ that was estimated from a training set (Eq. 5), misclassification of instance X occurs when the instance is "positive", but smaller than $\hat{A}$ or "negative" but greater than $\hat{A}$. Given a cost parameter of U and an estimated threshold of $\hat{A}$, the expected misclassification cost at is:

$$\begin{aligned} &C^C\left(\hat{A}\right) = U * \left(P\left(X < \hat{A} \middle| X \in Y_1\right) + P\left(X > \hat{A} \middle| X \in Y_0\right)\right) = \\ &U * \left(\Phi\left(\left(\hat{A} - \mu_1\right)/\sigma\right) + 1 - \Phi\left(\left(\hat{A} - \mu_0\right)/\sigma\right)\right) = \\ &U * \left(\Phi\left(\left(\hat{A} - \mu_1\right)/\sigma\right) + \Phi\left(\left(\mu_0 - \hat{A}\right)/\sigma\right)\right) \end{aligned} \quad (11)$$

C^C is minimized when $\hat{A} = A = 0.5*(\mu_0+\mu_1)$ (i.e., with a sample size $N \rightarrow \infty$):

$$C^{C*} = C^{C}\left(\hat{A} = A = 0.5*(\mu_0 + \mu_1)\right) = 2U*\Phi((\mu_1 - \mu_0)/2\sigma) \quad (12)$$

Given a finite sample-size N and a quality level Q^D (i.e., an actual sample size of Q^D*N) – we define the expected classification cost C^c as the mean of $C^c(\hat{A})$ for all possible values of the estimated threshold $\hat{A}$.

$$C^C = E\left[C^C\left(\hat{A}\right)\right] = U*E\left[\Phi\left(\left(\hat{A} - \mu_1\right)/\sigma\right) + \Phi\left(\left(\mu_0 - \hat{A}\right)/\sigma\right)\right] \quad (13)$$

The mean expression above relates to a confidence interval $L_A\left(1-\rho, N, Q^D\right)$ around the estimation $\hat{A}$, that includes the optimal threshold A (Eq. 3) given an actual sample size of Q^D*N, with a confidence rate of 1-□. Accordingly the expected classification cost can be defined as:

$$C^C\left(1-\rho, N, Q^D\right) = E\left[C^C\left(\hat{A}\right)\middle|\hat{A} \subset CI\right] = U*\frac{\int_{\hat{A}\subset CI}\left(\Phi\left(\left(\hat{A}-\mu_1\right)/\sigma\right) + \Phi\left(\left(\mu_0-\hat{A}\right)/\sigma\right)\right)d\hat{A}}{(1-\rho)*L_A\left(1-\rho, N, Q^D\right)} = U*\delta\left(1-\rho, N, Q^D\right) \quad (14)$$

$$where \quad CI = \left[\hat{A} - 0.5*L_A\left(1-\rho, N, Q^D\right), \hat{A} + 0.5*L_A\left(1-\rho, N, Q^D\right)\right]$$

The expression $\delta(1\text{-}\rho, N, Q^D)$ above reflects the average misclassification likelihood for a certain item, given certain values of confidence level 1-ρ, training-set size N, and DQ level Q^D. It is likely to decrease with a smaller ρ, larger N, and/or larger Q^D.

We may have the ability to complete missing values in our training set, at a cost of *S* units per missing items. The benefits gained from completing those values would justify the associated cost if the reduction in misclassification cost will be higher than the cost of missing-values completion. Assume that the current quality level is $Q^{D/S}$, and the target quality level is $Q^{D/T}$. If we have N^T items that need to be classified, the classification costs that will be saved by filling in missing values will be

$$\Delta C^C\left(Q^{D/T}\right) = N^T * \left(C^C\left(1-\rho, N, Q^{D/S}\right) - C^C\left(1-\rho, N, Q^{D/T}\right)\right) = N^T * U * \left(\delta\left(1-\rho, N, Q^{D/S}\right) - \delta\left(1-\rho, N, Q^{D/T}\right)\right) \quad (15)$$

The correction cost ΔC^S of increasing quality from $Q^{D/S}$ to a target level of $Q^{D/T}$ is:

$$\Delta C^S\left(Q^{D/T}\right) = S * N * \left(Q^{D/T} - Q^{D/S}\right) \quad (16)$$

Figure 4 highlights the interplay between the costs of correction versus misclassification. If the target is to reduce misclassification costs by ΔC^C, this will involve a correction cost of ΔC^S due to the need to complete missing values and raise the quality Q^D to the desired level. The net-benefit associating with missing-value completion is given by $B(Q^{D/T}) = \Delta C^C(Q^{D/T}) - \Delta C^S(Q^{D/T})$. We can now frame the question of setting the desired quality-level to a certain target as an optimization problem - choose $Q^{D/T}$ that maximizes:

$$B\left(Q^{D/T}\right) = N^T * U * \left(\delta\left(\rho, N, Q^{D/S}\right) - \delta\left(\rho, N, Q^{D/T}\right)\right) - S * N * \left(Q^{D/T} - Q^{D/S}\right) \quad (17)$$

S.t., $Q^{D/S} \le Q^{D/T} \le 1, B \ge 0$

Where,

B	The net-benefit associated with data quality improvement
$Q^{D/S}$, $Q^{D/T}$ -	The given data quality level and the target, respectively
1-ρ -	The target confidence level
N -	The number of samples in the complete training dataset
N^T -	The number of samples to be classified
δ (ρ, N, Q^D) -	The average likelihood of misclassification
U-	The expected cost of misclassifying a single item
S-	The cost of fixing a single missing value

The objective-function above is not-linear and has no close-form solution. However, the optimal solution can be approximated with a software-based optimizer. DQ management decisions often involve substantial cost-benefit tradeoffs (Ballou et al. 1998; Heinrich et al. 2009; Even et al. 2010). The need for cost-benefit assessment is also reflected in this study – but with differentiation between the datasets on which we act. The data correction cost is associated with the training set, used for building the model. On the other hand, the reduction in misclassification cost is associated with data items that are not part of the training set, but have to be classified according to the model developed.

2.4 Multivariate extension

We now adapt the model for the case of a multidimensional input vector with M attributes $X\{x_1,x_2,\ldots,x_M\}$. A linear classification rule in this case would be expressed as a hyperplane $A = f(x_1, x_2, \ldots, x_M) = a_0 + a_1 * x_1 + a_2 * x_2 + \cdots + a_M * x_M$. As with the scalar case, the

estimation of the coefficients vector $Z\{a_1, a_2, \dots, a_M\}$ is less precise with a smaller dataset. With scalar classification the estimation precision was conceptualized as a Confidence Interval (CI). Similarly, with multivariate classification, we define a multidimensional cubic Confidence Region (CR) that represents the extent of certainty in the estimation of the coefficients set Z. Given a set of coefficients Z, the CR can be defined as a set of CI $Z_{\mathrm{CR}} = \{CI_0, CI_1, CI_2, .., CI_M\}$, each representing the CI of the associated coefficient:

$$CI_i\left(\rho, Q^D, N\right) = \left[\hat{a}_i - t_{1-\rho/2,\lfloor 2*Q^D*N \rfloor - 2} * \hat{\sigma}_{\hat{a}_i},\quad \hat{a}_i + t_{1-\rho/2,\lfloor 2*Q^D*N \rfloor - 2} * \hat{\sigma}_{\hat{a}_i}\right] \tag{18}$$

Where,

$\hat{a}_i$ - The estimation of the coefficient a_i

$\hat{\sigma}_{\hat{a}_i}$ - The estimated standard deviation of $\hat{a}_i$

1-ρ - The target confidence level

N - The number of samples per class in the complete training dataset

Q^D - The data quality level (i.e., the rate of non-missing values)

$t_{1-\rho/2, Q^D N-2}$ The 1-ρ quantile of Student-t distribution with N degrees of freedom

Given this definition, the model quality is defined as a weighted product of the confidence intervals, considering a vector of weights $\omega_1, \omega_2, \dots, \omega_M$ that reflect the relative contribution of each variable:

$$Q_A^M\left(\rho,N,Q^D\right)=\prod_{i=0}^{m}\omega_i * CI_i\left(\rho,Q^D,N\right)=\Delta\left(\rho,N,Q^D\right) \quad (19)$$

To further illustrate the multivariate case, we consider the case of two bivariate normal distributions $f_0(x|\underline{\mu}_0,\sum)$ and $f_1(x|\underline{\mu}_1,\sum)$, overlapping to an extent, where the covariance matrix is diagonal (Figure 4).

Similarly to the scalar case, misclassification in the bivariate case can be defined in terms of false positive (FP) and false negative (FN) probabilities. It can be shown (Figure 5) that the optimal classifier, which minimizes the probability of FP+FN can be defined by a separation line. In a case where the line is unknown in advance but has to be calculated from a sample, it is possible to define a confidence region (CR) within which the optimal threshold resides with a likelihood of 1-ρ *I* (Figure 6). As with the scalar case, we sum up the misclassification costs for all possible decision rules within the CR. We accumulate the volumes of the probability density functions (PDF`s) that represent error rates (i.e., FP and FN), bounded by every possible decision rule that resided within the CR, and multiply it by the misclassification cost *U*:

$$C^C\left(\rho, N, Q^D\right) = E\left[C^C\left(\hat{A}\right)\middle|\hat{A} \subset CI\right] =$$

$$U * \frac{\int\int \ldots\ldots \int_{\hat{A} \subset \Delta} \left(F^* + G^*\right) da_1 da_2}{(1-\rho) * \Delta\left(\rho, N, Q^D\right)} = U * \delta\left(\rho, N, Q^D\right)$$

where

$$F^* = \int_{-\infty}^{\infty} \int_{a_0 + a_1 x_1}^{\infty} f_0(x \mid \mu_0, \Sigma) dx_2 dx_1 \qquad (20)$$

$$G^* = \int_{-\infty}^{\infty} \int_{\infty}^{a_0 + a_1 x_1} f_1(x \mid \mu_1, \Sigma) dx_2 dx_1$$

The decision cost represents a sum of all possible misclassification costs (i.e., misclassification probabilities multiplied by relevant costs), normalized by the confidence region of the classification rule coefficients.

3. EVALUATION

This section presents evaluation of the proposed framework with simulated and real-world data. Simulation, as a research method, has been used often in DQ management and Data Mining works (e.g., Blake et al., 2011, Askira-Gelman, 2011, Lauría. et al., 2011), for confirming theoretical hypotheses and/or for evaluating models and methodologies. The simulated experiments, described next, were executed with the MATLAB software package.

3.1 Experiment A: Evaluation with Simulated Scalar Data

The first experiment was conducted with simulated scalar (one dimensional) data. The samples were taken from two normally-distributed populations with means of μ_0=2, μ_1=4 and the same standard deviation of σ=3. The two classes had equal a-priory probabilities V_0=V_1=1/2 and equal misclassification costs. The experiment was conducted for a few sample sizes N = {200, 500, 1000, 10000}. For each sample size, the first run included the full dataset (Q^D =1), and then the data quality was gradually deteriorated by taking a smaller sample size, with Q^D<1. The DQ deterioration was achieved by random deletion of datasets records, with a mechanism that simulated a "missing completely at random" (MCAR) pattern. The model quality Q^M (the CI), as a function of the sample size, was then calculated (Eq. 7). Figure 6 shows the model quality (Q^M – the confidence interval length) versus data quality (Q^D) for different sample sizes, given ρ = 0.05.

The results (Figure 7) support our earlier arguments – the model quality, in terms of confidence interval, increases with a higher N and with a higher DQ level. Notably, with the highest sample-size shown (N=10000), the Q^M degradation is relatively minor for small Q^D degradation (Q^M (Q^D=1) = 0.08, versus Q^M (Q^D=0.6) = 0.1), but it becomes steeper as Q^D reaches low rates (Q^M (Q^D=0.1) = 0.26). It can be shown that with a large N, the Student-t distribution can be approximated by a Normal distribution - e.g., with 30 of more degrees of freedom, the error of approximating the PDF of a Student-t distribution with a Normal distribution is less than 0.005. Accordingly, the CI-length will be approximated by $L_A(\rho) = 2 * Z_{1-\rho/2}\hat{\sigma}_{\hat{A}}$

The classification cost (C^C) was expressed as a function of data quality (D^Q), based on Eq. 14, where the averaged variance of two partial samples (i.e., $Q^D * N_i$) was calculated per data quality level Q^D. The distribution parameters for the classification rule and the confidence interval were calculated for sample sizes N = {200, 500, 1000, 10000} and for different data quality levels between 0 and 1. Figure 8 shows the expected classification cost (C^C) versus the data quality (Q^D) for different sample sizes, with U=1 and ρ=0.05.

The similarity in behavior between Figure 7 and Figure 8 is noticeable – the expected cost is higher with lower sample size, and decreases further as the rate of missing values increases (lower Q^D). With the largest N, and with no missing values (Q^D=1), the expected C^C nearly reaches the optimum ($C^{C*} \approx 0.039$). At this large sample size the impact of missing values is relatively minor – a significant change in C^C can be noticed only when Q^D goes below 0.1.

3.2 Experiment B: Evaluation with Simulated Multivariate Data

The second experiment uses data samples that are drown from the multivariate, four-dimensional distributions (Marill and Green, 1963), which are commonly used for simulation studies. The mean vectors M_0 and M_1 as well as shared covariance matrix $\sum$ are given by:

$$M_0 = \begin{bmatrix} 7.825 \\ 6.75 \\ 8.525 \\ 7.065 \end{bmatrix} \quad M_1 = \begin{bmatrix} 5.76 \\ 5.715 \\ 4.15 \\ 6.96 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1.034 & 1.281 & -0.293 & 0.301 \\ 1.281 & 1.967 & -0.219 & 0.556 \\ -0.293 & -0.219 & 2.269 & 0.146 \\ 0.301 & 0.556 & 0.146 & 2.941 \end{bmatrix}$$

The four simulations were conducted for one to four dimensions, respectively. For J dimensions, the parameters are denoted by $M_0 = [m_{01} \ldots m_{0j}]$ and $M_1 = [m_{11} \ldots m_{1j}]$, with a shared

covariance matrix with diagonal components of $\{\sigma^2_1..\sigma^2_j\}$ and covariance components denoted by $\{\text{cov}(x_1x_2)\ldots\text{cov}(x_ix_j)\}$, where $i<j$ and $\{x_1\ldots x_j\}$ reflect dataset features. The model quality was calculated as a product of the confidence intervals (Eq. 19) for the discriminant coefficients, and was normalized to a 0-1 range.

Figure 9 reflects the 4-dimensional case, showing a similar behavior to Figure 7. The model quality is at maximum (lowest value) at perfect data quality, and shows exponential decay as data quality decreases. The model quality degradation appears to be more gradual in the multivariate case than in the scalar cases. This behavior is consistent for all sample sizes, but as the sample size grows, the Q^M degradation appears to be relatively small for high data quality, but gets steeper as data quality lowers. For example, at N=6,000 there is a 0.18 in degradation in Q^M between Q^D=1 and Q^D=0.6, versus 0.31 degradation in Q^M between Q^D=0.6 and Q^D=0.2.

The impact of decision quality on the classification cost (Figure 10) was evaluated by using the numeric approximation in Eq. 20. Initially, for each data quality level and sample size, we conducted multiple replications of linear discriminant analysis and calculated the confidence intervals. This was followed by conducting a numeric approximation of Eq. 20, by calculating the sum of error rates within a confidence region defined by CI of linear discriminant function coefficients. Similarly to the previous cases, Figure 10 shows a nearly-exponential decay the classification cost versus decision quality. Despite the presence of some noise, it seems that the value of classification cost is approximately the same for all quality level rates, represented by the product of full sample size and DQ rate Q^D*N (i.e., $C^C(Q^D=0.5, N=6{,}000) \approx C^C(Q^D=1, N=3{,}000)$).

3.3 Experiment C: Evaluation with Real-World Multivariate Data

The third experiment used a dataset received from a large bank. The dataset contained a random sample of 100,000 bank accounts, one record per account. Each record has a unique customer identifier and a few attributes that reflect financial characteristics and activities. Any identifying details were removed in order to protect privacy and confidentiality.

The evaluation looked at a classification of customers to those who registered to a new service offered by the bank versus those who did not (reflected in a "Registration" output variable with 1 or 0 values, respectively). Four input variables were considered – three that reflect the customer's financial status (*Income*, *Savings*, and *Debts*), and one that reflected the customer's Age. Figure 11 shows the distribution of those 4 variables. The distributions are asymmetrical - with *Income*, *Savings* and *Debts* the histograms reveal high concentration of low values with a long right tail. The distribution of Age is also asymmetrical and right-tailed, but here the high concentration is not in the low end, but rather around the ages of 30.

The training set included randomly-chosen 10000 records, balanced between the two output groups. The other records were used to assess the classification quality, following similar experiment procedures to those described in the first two experiments. The results were similar in nature to those gained in the previous experiments and reflect improvement in model quality (smaller confidence interval) with high completeness rate, following an approximately exponential decline pattern. Figure 12 shows the misclassification cost versus the data quality level for the third experiment. As with the previous experiment, the decision quality improves (lower misclassification cost) with higher data quality level at all sample sizes. The curves show

relatively mild decline, and only at very low DQ levels the decline becomes steeper. Notably, the behavior in this case is similar to the behavior that was observed in previous evaluations - even though the LDA assumptions are not met in this real-world dataset. The similarity in results may indicate robustness of the purposed methodology, even in cases where the underlying assumptions do not fully met.

To further explore the robustness of our method, we simulated CAR data degradation with few real-world datasets, which are often used in investigations of missing data (e.g., Luengo et al. 2010). The selected datasets *Bands* (N=539), *Adult* (N=48842), and *Marketing* (N=8993), (Alcalá-Fdez et al. 2011), have different ratios of missing values (32.28%, 7.41%, and 23.54% respectively) and can be used to train linear classifier. For each dataset we then initiated different missing value rates and calculated the proposed model and decision quality metrics. We varied the number of the explanatory variables between two to five and averaged the results. As expected, the results are very similar in nature to the previously outlined experiments and follow the exponential deterioration of model quality as a function of data quality (Figure 13). Additionally, the decision quality is exponentially decreasing with the deterioration of data quality, in consistency with earlier results (Figure 11).

To evaluate the robustness of our methodology, we performed cross-validation of the model and the decision quality metrics (Figure 14). While for some datasets the cross-validation shows nearly-identical results to the initial model quality, for others it shows some divergence. This divergence is mainly a result of the noise inherited in dataset and of the sample size. Notably, even when the curves representing model quality divert, the relationship between data and model

quality is consistent – exponential decline in model quality as the data quality linearly deteriorates.

4. CONCLUSION

The negative impact of DQ on decision making has been broadly acknowledged in research and in practice. This study suggests that a possible way for understanding and quantifying this impact is by looking into the mediating role played by decision-support models in classification problems. Such models are often estimated from training datasets – and when such a training dataset suffers from DQ defects, the model and the decisions that it supports are likely to be biased. This claim makes intuitive sense – however, not much was done so far to support it analytically. This study takes a step in that direction by offering an analytical framework that links the three levels of quality assessment - data quality, model quality, and decision quality. The analytical development aimed at demonstrating the key concepts of the proposed methodology, as a step towards fuller implementation in comprehensive and complex decision scenarios. The development and the evaluation applied a few simplifying assumptions at first, regarding symmetry conditions, and later relaxed some of these. Obviously, application of the proposed methodology under more realistic conditions would require more sophisticated analytical developments and numerical methods, which are ought to be explored in the future.

This study makes a number of contributions to Data and Information Quality research. It offers a theoretical conceptualization of three quality layers in typical decision scenarios - data, model and decision. The proposed approach highlights the mediating role of model quality and provides a quantitative viewpoint of the effect of missing data on constructed classification model and, subsequently on decision making. The proposed methodology highlights and

quantifies the relation between the quality aspects of three main layers – data, model, and decisions. The methodology uses relatively simple statistical means to demonstrate the formation mechanism of flawed model-based decision process. The study also presents an evaluation of the proposed methodology through simulations with both synthetic and real-world data. The findings highlighted an important behavior pattern that was common in all cases – exponential-like improvement in model and decision quality as data quality increases, with steep improvement in low levels of completeness and milder improvement at higher completeness rates.

Our study formulates the decision whether to complete the training dataset by acquiring missing data as a cost-benefit optimization problem. This formulation reflects a broader data management challenge - the need to balance between the costs associated with completing missing values in datasets toward increasing their quality level, versus the net-benefit associated with such completion. The optimization approach presented in this study, extends similar cost-benefit optimization concepts that were presented earlier (e.g., Ballou & Pazer, 1995, 2003; Even & Shankaranarayanan, 2007) by linking degradation in data quality to the resulting degradation in model and decision quality. Moreover, this study suggests directing the question of sufficient data quality level by the need to assure the quality of data-driven decisions. This implies that the question of data completeness must be interpreted not only through the lenses of missing-values count, but also as a question of having sufficient data to support efficient decisions and ensure their correctness. Knowing the level of confidence required by decision makers, who use certain dataset for classification and/or other decision tasks - a modeling approach such as the one presented in this study may help determining the optimal quality level of that dataset. This, by considering the characteristics of decision scenario, the efforts required to bring the dataset to a

certain quality level, the expected value gain by data quality improvement, and the associated cost-benefit tradeoffs.

Another practical contribution of this study lies in understanding the behavior of decision degradation dependently with quality of data (with respect to completeness dimension) as a key for developing efficient policies for DQ maintaining. Furthermore, we have proposed practical measurements of model and decision quality as a part of proposed methodology and their evaluation on various synthetic datasets and real-world data reveals to satisfying results in terms of robustness and consistency. Noteworthy, the results of simulation experiments highlight the lack of need in dramatic amount of data storage, to satisfy decisions with required level of trustworthiness. According to presented results, datasets with hundreds of tuples are sufficient in terms of decision quality, providing that the existing sample truly represents the real-world behavior. This conclusion, supported by statistical considerations, contradicts to arising tendency of collecting vast amounts of data in organizations` repositories, assuming that the more data are collected the better. Finally, the proposed analytical scope may be used as a basis for development of advanced, dynamic algorithms for DQ monitoring to improve data quality awareness. It may also help assessing the necessity of data imputations, and the magnitude of investments in imputation solutions that can be justified.

While making a few important contributions, the proposed methodology has some limitations, which future extensions can potentially address. The analytical development demonstrated in this study is relatively simple, as its aim was to highlight and demonstrate the key concepts. Fuller development is likely to require more advanced techniques that can potentially address more complex decision scenarios. The framework development has made a few limiting assumptions.

Some of this should be handled further, and possibly relaxed in future extensions. For example: a) Classification scenarios with more than two classes, b) Discrimination parameters that do not follow a normal distribution, c) Classification models other than the Linear Discriminant Analysis (e.g., Quadratic Discriminant Analysis, Nearest Neighbor algorithm), and d) Missing values that are not completely random, hence must be handled under a pattern assumption other than MCAR (e.g., MAR – Missing at Random; NMAR – Not missing at Random).

Other possible extensions can be considered. For instance, exploring other data quality dimensions (e.g., currency, accuracy), beyond the completeness dimension studied in this research. The criterion for model quality assessment was chosen to be the confidence interval (CI), which was calculated for each parameter of the classification rule. The CI was chosen for being commonly used and relatively simple to calculate. However, an alternative approach could have been to develop a criterion that considers the quality for the entire model, and not only for specific parameters. The goodness of decision measured by means of minimizing the classification cost while maximizing accuracy, precision, sensitivity, and/or specificity may be considered as optional for alternative criteria for decision quality assessment as well. Lastly, the decision scenario that was assessed is classification – however the question of data, model, and decision quality arises in other scenarios (e.g., optimization within a continuous range), and worth further exploration as well.

To conclude, this study presents analytical framework that links between data, model and decision quality. The proposed approach may contribute to understanding the degradation of decision performance when the quality of data is damaged. The presented methodology can be used to find breakeven point that optimizes the balance between the correction costs, aimed to

increase the quality level of dataset, and the net-benefit revenues associated with missing-value competition. Finally, the proposed analytical scope may be used as a basis for development of advanced, dynamic algorithms for DQ monitoring.

REFERENCES

Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F. (2010). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, *17*(2-3), 255-287.

Askira Gelman, I. (2011). GIGO or not GIGO: The Accuracy of Multi-Criteria Satisficing Decisions. *Journal of Data and Information Quality (JDIQ) 2, 9.*

Ballou, D. P., Pazer, H. L. (2003). Modeling completeness versus consistency tradeoffs in information decision contexts. *Knowledge and Data Engineering, IEEE Transactions On, 15*(1), 240-243.

Ballou, D. P., Pazer, H. L. (1995). Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research, 6*(1), 51-72.

Ballou, D., Wang, R., Pazer, H., Tayi, G.K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science* 44, 462-484.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)* 41, 16.

Bhargava, H.K., Power, D.J., Sun, D. (2007). Progress in Web-based decision support technologies. *Decision Support Systems* 43, 1083-1095.

Blake, R., Mangiameli, P. (2011). The effects and interactions of data quality and problem complexity on classification. *Journal of Data and Information Quality (JDIQ)* 2, 8.

Brown, M. L., Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems, 103*(8), 611-621.

Duda, R.O., Hart, P.E., Stork, D.G. (2012). *Pattern classification*. Wiley-interscience.

Even, A., Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *ACM SIGMIS Database* 38, 75-93.

Even, A., Shankaranarayanan, G., Berger, P. D.: Economics-Driven Design for Data Management: An Application to the Design of Tabular Datasets. *IEEE Tr. on Knowledge and Data Engineering* (19:6), 818-831.

Even, A., Shankaranarayanan, G. and Berger, P.D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting. *Decision Support Systems* 50, 152-163.

García-Laencina, P. J., Sancho-Gómez, J., Figueiras-Vidal, A. R., Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*(7), 1483-1493.

Hastie, T., Rosset, S., Tibshirani, R., Zhu, J. (2004). The entire regularization path for the support vector machine. *Ann Arbor*, *1001*, 48109-1092.

Heinrich, B., Klier, M., Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data and Information Quality (JDIQ)* 1, 5.

Lauría, E. J., March, A. D. (2011). Combining bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis. *Journal of Data and Information Quality (JDIQ), 2*(3), 13.

Luengo, J., García, S., Herrera, F. (2010). A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method. *Neural Networks*, *23*(3), 406-418.

Luengo, J., García, S., Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, *32*(1), 77-108.

Kotsiantis, S., Koumanakos, E., Tzelepis, D., Tampakas, V. (2007). Forecasting Fraudulent Financial Statements using Data Mining. *International Journal of Computational Intelligence*, 3(2).

Madnick, S.E., Wang, R.Y., Lee, Y.W., Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)* 1, 2.

Marill, T., Green, D. (1963). On the effectiveness of receptors in recognition systems. *Information Theory, IEEE Transactions On, 9*(1), 11-17.

Mclachlan, G.J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley-Interscience.

Oehlert, G.W. (1992). A note on the delta method. *The American Statistician* 46, 27-29.

Pipino, L.L., Lee, Y.W., Wang, R.Y. (2002). Data quality assessment. *Communications of the ACM* 45, 211-218. .

Redman, T.C. (1996). *Data Quality for the Information Age*, Artech House, Boston.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems* 33, 111-126.

Van Der Heijden, F., Duin, R., De Ridder, D., Tax, D. M. (2005). *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. John Wiley & Sons.

Webb, A. R. (2003). Statistical pattern recognition. *Wiley*.

Table 1: Binary classification assessment with a 2-Way confusion matrix

Real-World Class	**Classification**	
	1	**0**
1	**True Positive (TP):** Correctly classified positive instances	**False Negative (FN):** Positive instances, incorrectly classified as negative
0	**False Positive (FP):** Positive instances, incorrectly classified as negative	**True Negative (TN):** Correctly classified negative instances

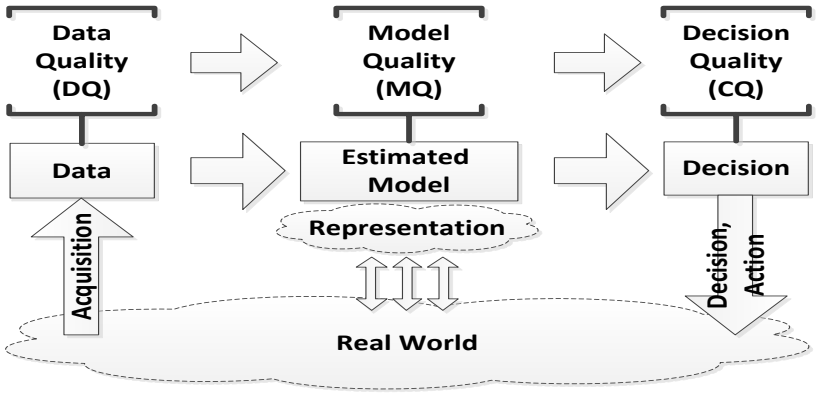

Fig. 1. A decision process

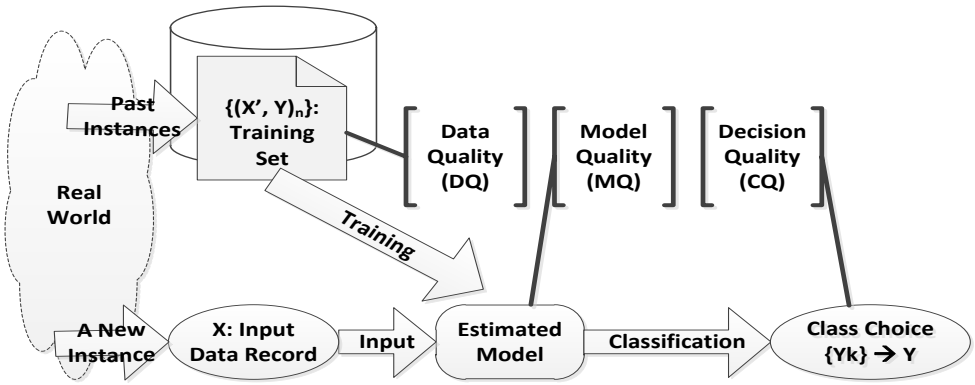

Fig. 2. The general methodology

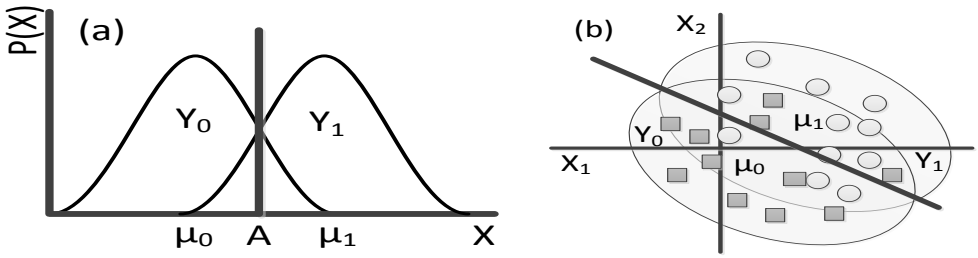

Figure 3. LDA Classifiers for (a) 1-dimensioal space, and (b) 2-dimensional space

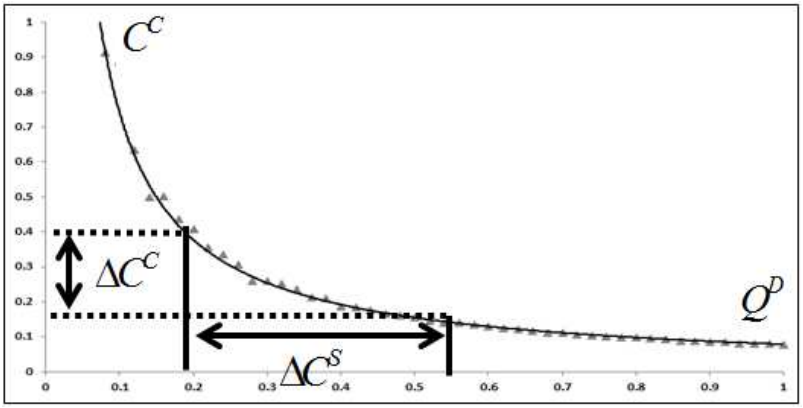

Figure 4. Reduction in Classification Cost versus Improvement in Data Quality

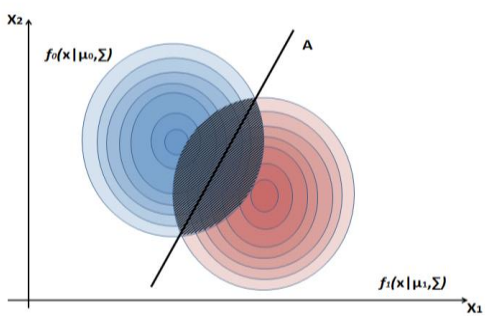

Figure 5. Two overlapping bivariate normal distributions

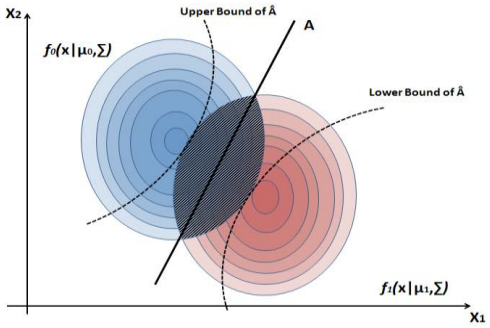

Figure 6. The confidence range of the classification hyperplane

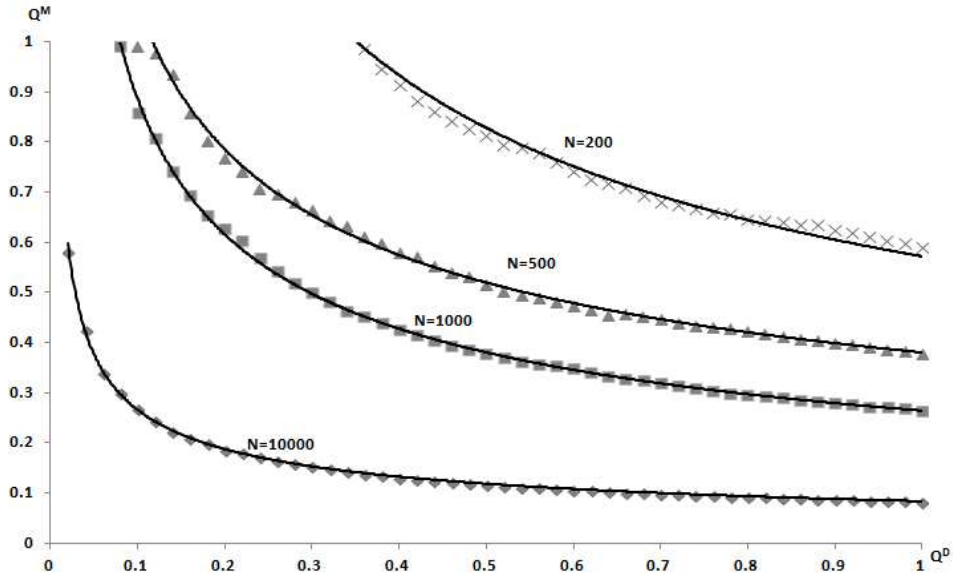

Figure 7. Experiment A - Model Quality versus Data Quality

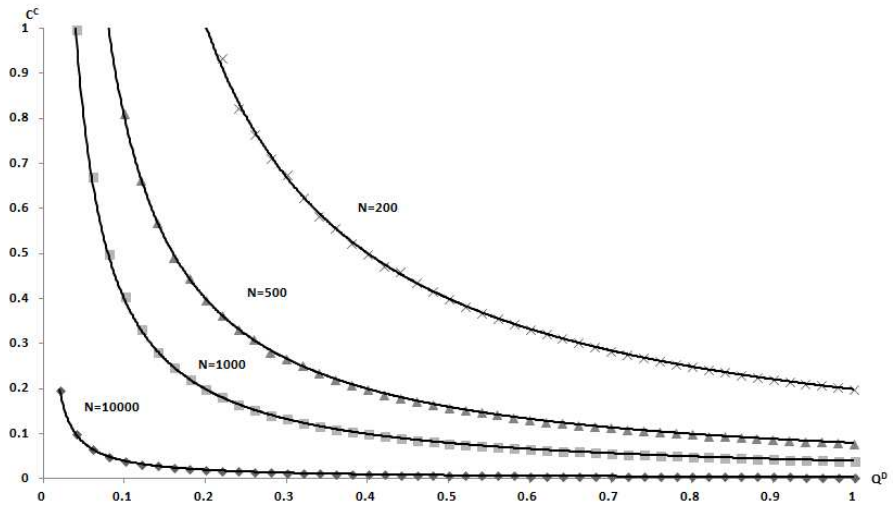

Figure 8. Experiment A - Decision Quality versus Data Quality

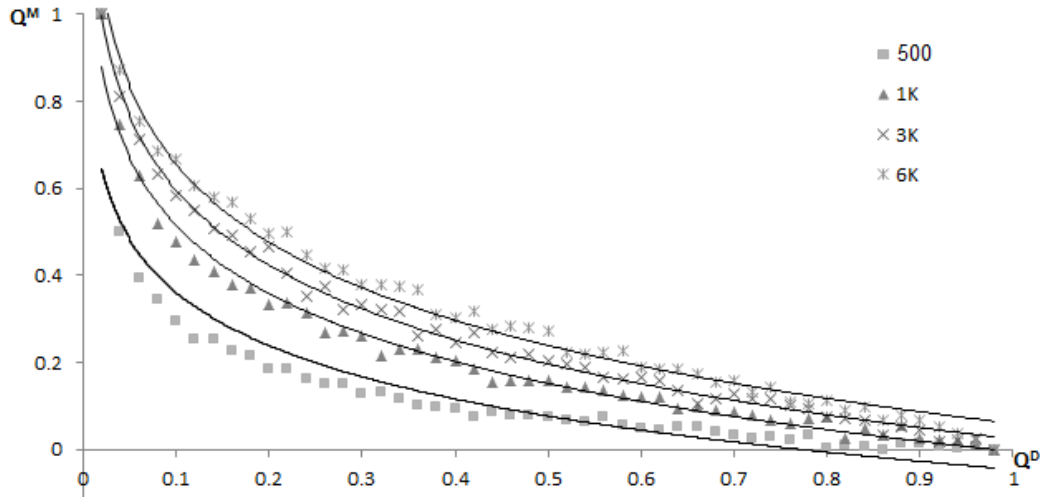

Figure 9. Experiment B - Model Quality versus Data Quality

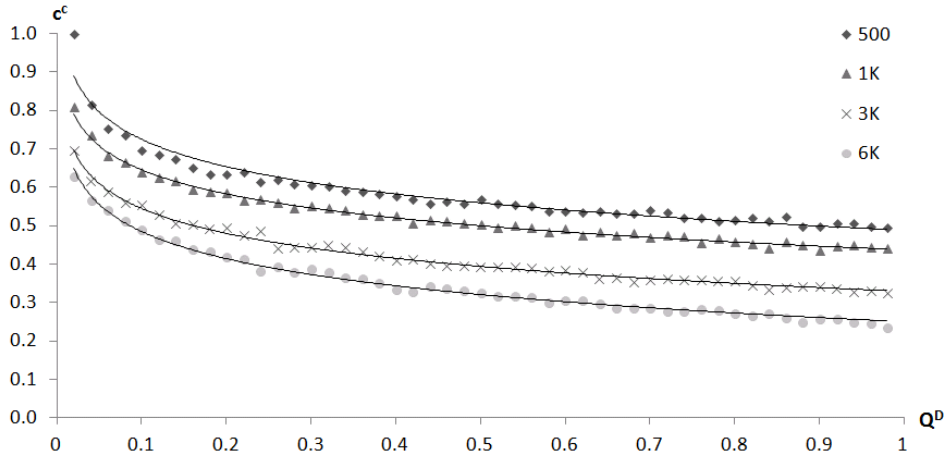

Figure 10. Experiment B - Decision Quality versus Data Quality

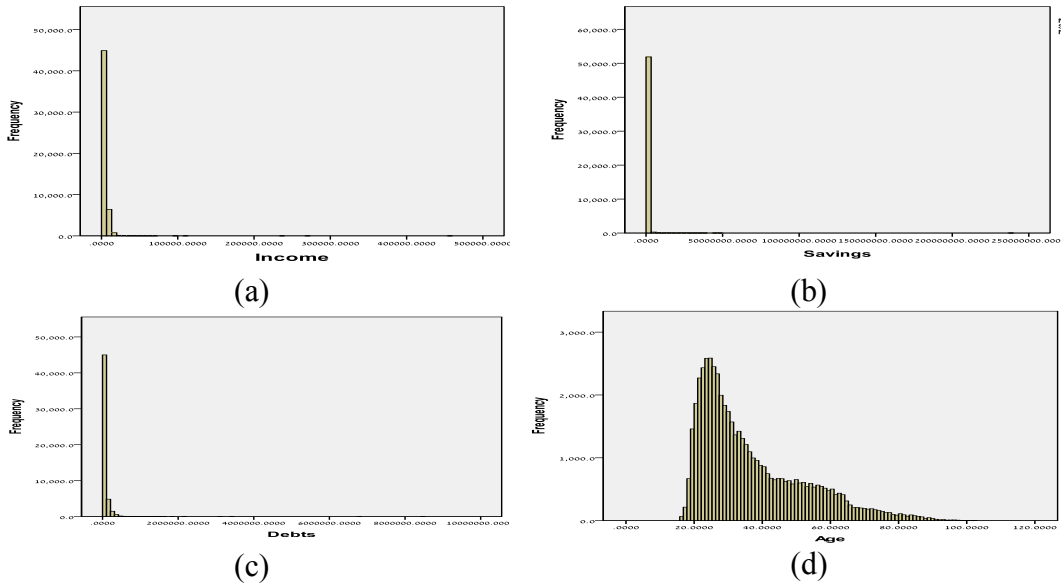

Figure 11. Experiment C - Model Quality versus Data Quality

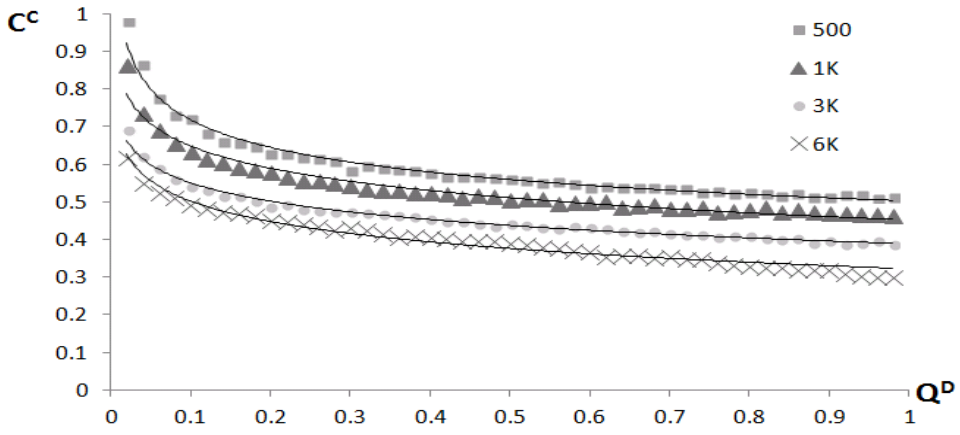

Figure 12. Experiment C - Decision Quality versus Data Quality

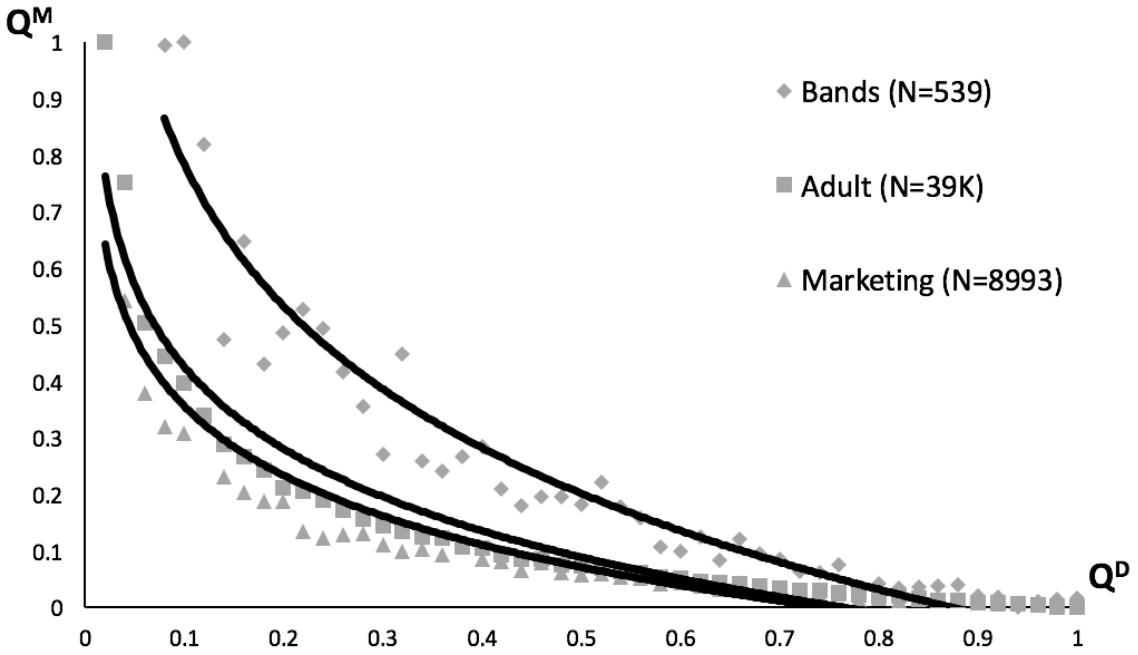

Figure 13. Evaluation of Model Quality versus Data Quality with three real-world datasets

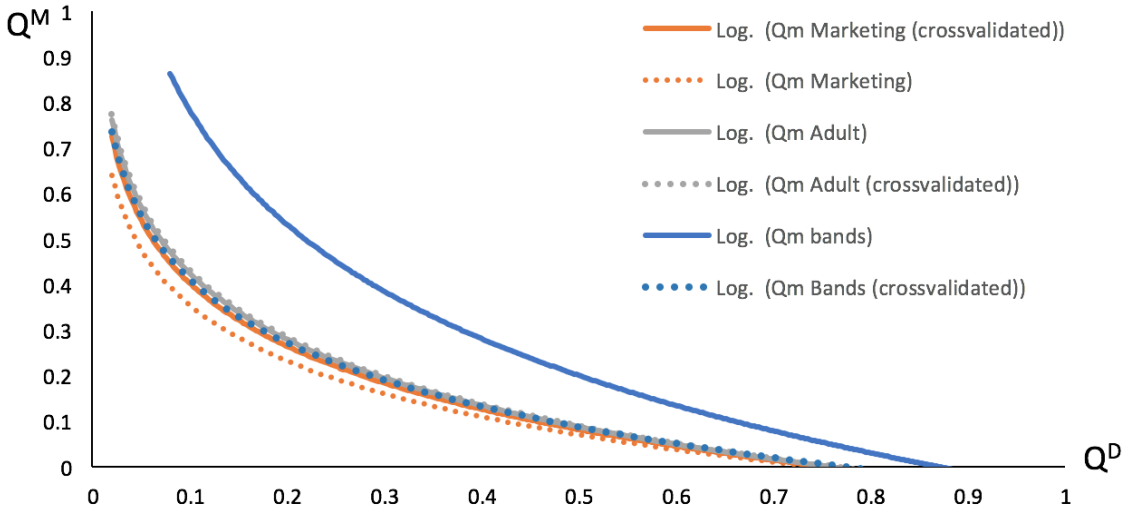

Figure 14. Evaluation of Model Quality versus Data Quality with three real-world datasets – a comparison of cross-validated and non cross-validated results